Equality for worst-case work at any protocol speed

Oscar C. O. Dahlsten,[1] Mahn-Soo Choi,[2] Daniel Braun,[3]
Andrew J. P. Garner,[1] Nicole Yunger Halpern,[4] and Vlatko Vedral[1, 5]

[1] *Atomic and Laser Physics, Clarendon Laboratory,*
University of Oxford, Parks Road, Oxford OX13PU, United Kingdom
[2] *Department of Physics, Korea University, Seoul 136-701, South Korea*
[3] *Institut für theoretische Physik, Universität Tübingen,*
Auf der Morgenstelle 14, 72076 Tübingen, Germany
[4] *Institute for Quantum Information and Matter, Caltech, Pasadena, CA 91125, USA*
[5] *Center for Quantum Technologies, National University of Singapore, Republic of Singapore*

(Dated: April 21, 2015)

We derive an equality for non-equilibrium statistical mechanics. The equality concerns the worst-case work output of a time-dependent Hamiltonian protocol in the presence of a Markovian heat bath. It has the form "worst-case work = penalty - optimum." The equality holds for all rates of changing the Hamiltonian and can be used to derive the optimum by setting the penalty to 0. The optimum term contains the max entropy of the initial state, rather than the von Neumann entropy, thus recovering recent results from single-shot statistical mechanics. We apply the equality to an electron box.

General Introduction—Average values of quantities are not always typical values: Outcomes may fluctuate significantly. In non-equilibrium nano and quantum systems this is often the case, with, for example, the work output of a protocol having a significant probability of deviating from the average. Hence, in these important systems, statements about averages have limited use when it comes to predicting what will happen in any given trial; the fluctuations need to be discussed explicitly.

Two key relations concerning fluctuations in work, Crooks' Theorem [1] and Jarzynski's Equality [2], have been studied extensively theoretically and experimentally. Amongst other things they can be used to determine free energies of equilibrium states from non-equilibrium experiments.

A recently developed alternative approach to non-equilibrium statistical mechanics is *single-shot statistical mechanics* [3–12], inspired by single-shot information theory [13, 14]. The focus is on statements that are guaranteed to be true in every trial, rather than on average behaviors. For example, one can ask whether a process's work output is guaranteed to exceed some threshold value (such as an activation energy), or whether a process's work cost is guaranteed not to exceed some threshold value (beyond which the system may break from dissipating heat). These statements concern the *worst-case work* of a process. A key realisation is that the optimal worst-case work is determined not by the von Neumann/Shannon entropy of the initial state, but rather the max entropy, which is the logarithm of the number of non-zero eigenvalues of the density matrix. Thus, which entropy one should use in statements about optimal work depends on which property of the work probability distribution one is interested in.

Single-shot statistical mechanics began with almost no *a priori* relation to fluctuation theorems, but promising links were made in [6, 15]. We shall use two realizations from [15], namely that (i) in the trajectories model for work extraction, both single-shot and fluctuation results apply; and (ii) Crooks' Theorem can be used to make a certain statement about worst-case work. A natural question that arose from these results is how to link Crooks' Theorem to the existing single-shot statements concerning optimal work in terms of the entropy of the initial state.

We here show that key expressions concerning optimal worst-case work from [3, 5, 6] follow from Crooks' Theorem plus some extra thought. We moreover generalise them by giving an equality for the worst-case work that holds for any protocol in the set-up, including fast protocols. The equality holds in every process in a general set-up that involves a time-varying Hamiltonian and a single Markovian heat bath, modelled using trajectories. It has the form 'worst-case work=penalty-optimum,' and the optimum can thus be derived by setting the penalty to zero. To make the link to physics clear, we apply the result to an electron box experiment [16–18].

We begin with defining the set-up.

One-shot relative entropies—The standard relative entropy is $D(\rho||\sigma) := -\mathrm{Tr}(\rho[\log\rho - \log\sigma])$ [19], where log in this paper means the natural logarithm also known as ln. This is part of a wider class of relative entropies known as the Renyi relative entropies, which are parameterized by an integer α. We shall use two other members of that family: the (classical version of the) ∞ relative entropy $D_\infty(P||Q) := \sup_x \log(\frac{p_x}{q_x})$ and the 0 relative entropy $D_0(\rho||\sigma) := -\mathrm{Tr}(\pi_\rho \log\sigma)$, wherein π_ρ projects onto the support of ρ [20]. These are called one-shot relative entropies as they arise naturally in one-shot (also called single-shot) information theory [13, 14, 20].

Protocols, trajectory model of—We now describe the theoretical model, using the notation of [21]. The physical scenario we have in mind is depicted in Fig. 1

A protocol will be a sequence of elementary changes: (i) changes of the Hamiltonian and (ii) thermalizations. We shall initially assume there is a finite number of such

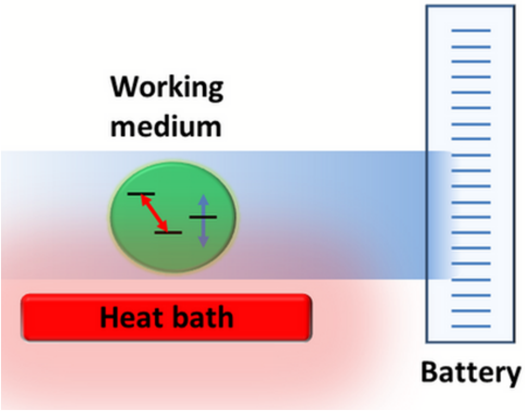

Figure 1: A schematic of the setup. There is a working-medium system, a battery system from which work is taken or given to, and a single heat bath. The battery system has the effect of altering the Hamiltonian of the working-medium, depicted with the blue arrow shifting an energy level. The heat bath has the effect of hopping the system between energy levels, depicted by the red arrow.

steps (but later show that the continuum limit is well-defined and corresponds to a master equation, at least in the discrete-classical case). The Hamiltonian is parameterized by λ_m, where m is an integer that labels the step.

1. Hamiltonian changes map λ_m to λ_{m+1}. We follow [21] in supposing there is an energy measurement in the instantaneous energy eigenbasis at the beginning and end of each Hamiltonian-changing step. In a given realisation the system then evolves from $|i_m, \lambda_m\rangle$ to $|i'_m, \lambda_{m+1}\rangle$, where i_m labels the energy eigenstate. This costs work given by the energy difference: $w_m = E(|i_m, \lambda_{m+1}\rangle) - E(|i'_m, \lambda_m\rangle)$. An important special case is $i_m = i'_m$, which arises in the quasi-static (quantum adiabatic) limit, as well as if the energy eigenbasis is constant and only the energy eigenvalues change; this can be termed the discrete-classical case.

2. Thermalizations map i'_m to i_{m+1}, cost no work, and preserve the Hamiltonian: $|i'_m, \lambda_{m+1}\rangle \to |i_{m+1}, \lambda_{m+1}\rangle$. For notational simplicity let us label this as $|i\rangle \to |j\rangle$ with energy $E_i \to E_j$. The hopping probabilities respect thermal detailed balance: $\frac{p(|i\rangle \to |j\rangle)}{p(|j\rangle \to |i\rangle)} = e^{-\beta(E_j - E_i)}$. The energy change $E_j - E_i$ from such a step is called heat, Q_m.

A *trajectory* is the time-sequence of energy eigenstates occupied: $|i_0, \lambda_0\rangle \to |i'_0, \lambda_1\rangle \to |i_1, \lambda_1\rangle \to \ldots \to |i_{f-1}, \lambda_{f-1}\rangle \to |i'_{f-1}, \lambda_f\rangle \to |i_f, \lambda_f\rangle$.

The probability of a given trajectory is accordingly, assuming a Markovian heat bath,

$$p(traj) = p(|i_0, \lambda_0\rangle)\times \prod_{m=0}^{f} p(|i_m, \lambda_m\rangle \to |i'_m, \lambda_{m+1}\rangle)\times p(|i'_m, \lambda_{m+1}\rangle \to |i'_{m+1}, \lambda_{m+1}\rangle). \tag{1}$$

A trajectory's *inverse* is the reverse of the sequence. The inverse corresponds, in the discrete-classical case, to the Hamiltonian changes running in reverse, from λ_f to λ_0, and to the same thermalizations as in the forward protocol, with the sequence exactly inverted. This process is termed the reverse process. Beyond the discrete-classical case, the unitary associated with the inverse process Hamiltonian change is defined such that $p(|i'_m, \lambda_{m+1}\rangle \to |i_m, \lambda_m\rangle) = p(|i_m, \lambda_m\rangle \to |i'_m, \lambda_{m+1}\rangle)$. Our results will hold under that condition. There are at least two ways of satisfying that condition: (i) Simply let the unitary of the corresponding elementary step in the reverse process be U^{-1}, where U is that of the forwards process, (ii) apply a suitable 'time-reversal' operator Θ to all states and operators involved, as in [21]. The reverse trajectory is then the reverse sequence of the time-reversed energy eigenstates: $\Theta|i_f, \lambda_f\rangle ... \Theta|i_0, \lambda_0\rangle$, with the condition $p(|i'_m, \lambda_{m+1}\rangle \to |i_m, \lambda_m\rangle) = p(\Theta|i_m, \lambda_m\rangle \to \Theta|i'_m, \lambda_{m+1}\rangle)$ being satisfied, as time reversal implies taking the complex conjugate of the states, in a preferred basis, and the transpose of the time-evolution in the same basis: $U \to U^T$. The condition is thus satisfied as $\langle b|U|a\rangle = (\langle a|U^\dagger|b\rangle)^* = \langle a|^* U^T |b\rangle^*$.

A given trajectory has some work cost $w = \sum_m w_m$, in line with the definition of the Hamiltonian-changing steps. The inverse trajectory has work cost $-w$. A given protocol on a given initial state induces some probability distribution over trajectories, with an associated probability distribution over work $p(w)$. The forwards and reverse protocol gives rise to $p_{\text{fwd}}(w)$ and $p_{\text{rev}}(-w)$ respectively.

If the initial density matrix of the forwards process and reverse processes are both thermal, i.e. $\exp-(\beta H(\lambda_0))/Z_0$ and $\exp-(\beta H(\lambda_f))/Z_f$ respectively, Crooks' Theorem holds [21]:

$$\frac{p_{\text{fwd}}(w)}{p_{\text{rev}}(-w)} = \frac{Z_f}{Z_0}\exp(\beta w). \tag{2}$$

(To derive it take the ratio of Eq. 1 and the corresponding reverse trajectory expression. Apply thermal detailed balance and the equality of reverse hopping probabilities for the Hamiltonian-changing steps. Sum over trajectories with the same w, and note that the reverse of a trajectory has the same work up to a minus sign).

Worst-case work—The central object of interest is the *worst-case work*

$$w^0 := \max\{w : p(w) > 0\},$$

also known as the *guaranteed work* [7]. In practice this may be realised by some very unlikely trajectory, and it is then natural to consider the worst-case work of some subset of trajectories $\mathcal{T}$:

$$w^0_{\mathcal{T}} := \max\{w : p(w) > 0 \text{ and traj} \in \mathcal{T}\}.$$

Equality for worst-case work— Consider an initial state ρ_0, and a protocol of thermalizations and Hamiltonian changes with initial and final Hamiltonians $H(\lambda_0)$

and $H(\lambda_f)$ respectively. This induces a work probability distribution $p(w)$ and an associated w^0. We shall derive an equality of the form $w^0 =$ penalty - optimum.

We consider initial states of form $\rho_0 = \sum_i p_i |i_0, \lambda_0\rangle\langle i_0, \lambda_0|$, i.e. diagonal in the energy eigenbasis though not necessarily thermal (energy coherence may still arise during the protocol). We take $p_i \neq 0$. This is because we wish to avoid divergences from dividing by p_i. (See [22] for an alternative way of approaching this divergence problem).

To apply Crooks' Theorem (Eq. 2) here, even though the initial state is not assumed to be thermal, our approach is as follows. Note that if a state is not thermal, e.g. if one has a degenerate two-level system the thermal state is $\gamma = 1/2|0\rangle\langle 0| + 1/2|1\rangle\langle 1|$, but if one instead had $\rho_0 = 2/3|0\rangle\langle 0| + 1/3|1\rangle\langle 1|$, then this scenario has the same worst case work as γ. This follows because the set of trajectories with non-zero probability is the same in both cases, as can be seen from Eq.1 which gives the probability of a trajectory. Given a ρ_0 we will then find a corresponding thermal state which has the same worst-case work and apply Crooks' Theorem to that.

An important practical consideration which makes this more subtle is that some p_i may be negligible. It is then natural to exclude trajectories starting in those states when calculating the worst-case work. We therefore divide the initial energy eigenstates into two sets: one set which is the one of interest: $\mathcal{E}_{\mathrm{IN}}$ and the rest which we call $\mathcal{E}_{\mathrm{OUT}}$, corresponding to those we shall exclude when calculating the worst-case work. The probability of being in $\mathcal{E}_{\mathrm{OUT}}$ is given by

$$p(\mathrm{OUT}) = \sum_{|i_0,\lambda_0\rangle \in \mathcal{E}_{\mathrm{OUT}}} \mathrm{Tr}(|i_0, \lambda_0\rangle\langle i_0, \lambda_0|\rho_0).$$

We define $\mathcal{T}_{\mathrm{IN}}$ as the set of possible (meaning $p > 0$) trajectories beginning in $\mathcal{E}_{\mathrm{IN}}$ and similarly $\mathcal{T}_{\mathrm{OUT}}$ as the set of possible trajectories beginning in $\mathcal{E}_{\mathrm{OUT}}$. Recall that each trajectory has some work value associated with it. We call the worst-case work of $\mathcal{T}_{\mathrm{IN}}$, w^0_{IN}; this cannot be worse than the worst-case over all trajectories: $w^0_{\mathrm{IN}} \leq w^0$.

Now we design an associated thermal state to yield the same worst-case work as ρ_0, i.e. w^0_{IN} and later show this to be indeed be the case under an additional mild assumption. We define it as

$$\widetilde{\gamma} = \sum_{|i_0,\lambda_0\rangle \in \mathcal{E}_{\mathrm{IN}}} \frac{e^{-\beta E_i}}{\widetilde{Z}} |i_0, \lambda_0\rangle\langle i_0, \lambda_0| + \sum_{|i_0,\lambda_0\rangle \in \mathcal{E}_{\mathrm{OUT}}} p_i |i_0, \lambda_0\rangle\langle i_0, \lambda_0|,$$

changing the energies of $\mathcal{E}_{\mathrm{OUT}}$ to new ones, $\widetilde{E}_i$, such that $p_i = \exp(-\beta \widetilde{E}_i)/\widetilde{Z}$, and leaving the other energy levels the same. Our definition implies that

$$\widetilde{Z} = \frac{\sum_{|i_0,\lambda_0\rangle \in \mathcal{E}_{\mathrm{IN}}} e^{-\beta E_i}}{1 - p(OUT)}. \tag{3}$$

This partition function differs from that of the actual Hamiltonian $H(\lambda_0)$. Ignoring the $\mathcal{E}_{\mathrm{OUT}}$ levels helps lower the calculated work cost, as can be seen in the $\widetilde{Z}$ being smaller.

In this scenario with $\widetilde{\gamma}$ as the initial state and the $\mathcal{E}_{\mathrm{OUT}}$ levels lifted the protocol is the same as in the actual scenario, except that initially the $\mathcal{E}_{\mathrm{OUT}}$ are lowered down to the levels of the actual Hamiltonian of interest. The worst-case work of this scenario is called $\widetilde{w}^0$. We show (see Methods) that under a mild additional assumption that the worst-case work is bounded from below,

$$\widetilde{w}^0 = w^0_{\mathrm{IN}}, \tag{4}$$

as desired.

To get $\widetilde{w}^0$ from Crooks' Theorem (Eq. 2) we follow [15]. Take the initial state of the forwards process to be $\rho_0 = \widetilde{\gamma}$; and the initial state of the reverse process as $\gamma = e^{-\beta H(\lambda_f)}/Z_f$. Consider the equality of Crooks' Theorem (for values of w such that $p_{\mathrm{fwd}}(w) > 0$) and select the value for w which maximises the LHS (and thus the RHS) [15]:

$$\max \frac{\widetilde{p}_{\mathrm{fwd}}(w)}{\widetilde{p}_{\mathrm{rev}}(-w)} = \max \frac{Z_f}{\widetilde{Z}} e^{\beta w}.$$

The RHS is monotonic in w, so maximizing the RHS over the support of $\widetilde{p}_{\mathrm{fwd}}(w)$ leads to the maximum w-value w^0. Taking the logarithm and recalling the D_∞ definition yields [15]

$$\beta \widetilde{w}^0 = D_\infty(\widetilde{p}_{\mathrm{fwd}}(w)||\widetilde{p}_{\mathrm{rev}}(-w)) - \log\left(Z_f/\widetilde{Z}\right). \tag{5}$$

Main result— Combining Eq.5 and Eq. 4 we thus have

$$\beta w^0_{\mathrm{IN}} = D_\infty(\widetilde{p}_{\mathrm{fwd}}(w)||\widetilde{p}_{\mathrm{rev}}(-w)) - \log\left(Z_f/\widetilde{Z}\right). \tag{6}$$

Thus the worst case work of the trajectories of interest w^0_{IN} is this equal to (kT times) a relative entropy minus (the logarithm) of two partition functions, one of which encodes information about how many of the initial energy eigenstates have negligible occupation probability.

Discussion—Equation 6 has the form

$$\beta w^0 = \text{ penalty - optimum.}$$

The penalty is given by the difference between the forward and reverse distributions, quantified by D_∞. The optimum one can hope for, with a given initial state and given initial and final Hamiltonian, is to set the penalty to 0 (as relative entropies are non-negative), which leaves $-\log\left(Z_f/\widetilde{Z}\right)$. This term is made more negative the smaller the support of ρ is and the lower the final energies are relative to the initial ones. To illustrate the notation used, a very simple example of applying the formula is given in Fig. 2.

We now consider the optimum term in two important special cases where the single-shot entropy of the initial state emerges: (i) If $p(OUT) \to 0$ and $H(\lambda_0) = H(\lambda_f)$,

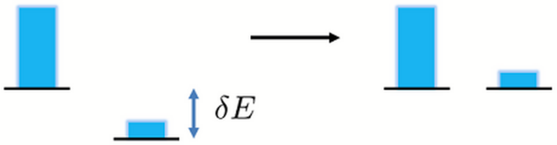

Figure 2: A very simple example of how to apply the formula. The energy levels are black lines and the occupation probabilities are indicated by the height of the blue rectangles. The forwards protocol here is to lift the second level from $-\delta E$ to 0. The reverse is to lower it back. Suppose for concreteness that $\rho_0 = 0.9|0,\lambda_0\rangle\langle 0,\lambda_0| + 0.1|1,\lambda_0\rangle\langle 1,\lambda_0|$.

This table describes the two possible trajectories, their work costs and probabilities (reverse trajectories in brackets).

	traj	traj set	work w	prob
traj 1	$|0,\lambda_0\rangle \underset{(\leftarrow)}{\rightarrow} |0,\lambda_f\rangle$	$\in \mathcal{T}_{IN}$	0 (0)	0.9 (0.5)
traj 2	$|1,\lambda_0\rangle \underset{(\leftarrow)}{\rightarrow} |1,\lambda_f\rangle$	$\in \mathcal{T}_{OUT}$	δE (-δE)	0.1 (0.5)

$\beta w^0_{IN} = 0$; $D_\infty = \log(2(0.9))$; $\log\left(\frac{Z_f}{\widetilde{Z}}\right) = \log(\frac{2}{1/0.9})$. Thus the equality is in this case: $0 = \log 2(0.9) - \log 2(0.9)$.

then $\log\left(Z_f/\widetilde{Z}\right) = -\log\sum_{i\in\mathrm{supp}(\rho_0)} e^{-\beta E_i}/Z_f = D_0(\rho_0||\gamma)$. Thus in this case the equality has the form

$$\beta w^0 = D_\infty - D_0.$$

(ii) If $p(OUT) \to 0$ and $H(\lambda_0) = H(\lambda_f) = 0$, $D_0(\rho_0||\gamma) = \log d - S_{\max}(\rho_0)$ (noting $\gamma = \mathbb{1}/d$ and recalling $S_{\max}(\rho) := S_0(\rho) := \log(\mathrm{rank}(\rho))$. This recovers the known results from [3, 5, 6] that these are optimal in the respective cases. (In the more general case where $p(OUT)$ is finite one recovers the *smooth* relative entropy as the optimal quantity–see Methods). The message is that it is the max entropy $S_{\max}$ which determines the optimal worst-case work, rather than the von Neumann entropy. If one defines thermodynamic entropy in terms of optimally extractable *worst-case* work, it is the max entropy which should be used.

To make the connection to physics clear, we apply the results to a recent realization of a Szilard engine with an electron box [16–18]. A great advantage with using this trajectories model from the fluctuation theorem approach is that it allows the application of single-shot results to such experiments. We described the set-up in in Figure 3 and in the Methods we analyse what controls the penalty term D_∞ in this scenario.

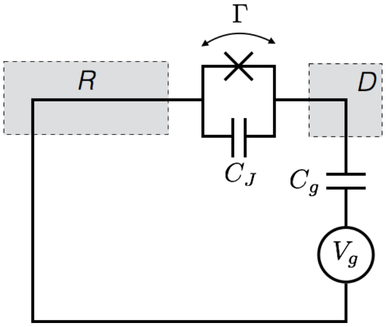

Figure 3: A schematic of an "electron box" (D) coupled to a metallic electrode (R) via tunnelling (with rate Γ) and the capacitor with capacitance C_J, and to the gate electrode via the capacitor with C_g. The gate voltage V_g controls the number of excess electrons on the electron box, which at low temperatures is restricted to two possible values and serves as a logical basis $|0\rangle$ and $|1\rangle$ for a qubit. Namely, it tunes the relative energy by $H \propto -C_g V_g |1\rangle\langle 1|$. The electrode R plays the role of a heat bath, where the tunnelling in/out of the box D corresponds to thermal excitation/relaxation. Experimentally, the work and heat can be measured by probing the charge on D in real time with a single-electron transistor next to D (not shown in the figure) as demonstrated in Refs. [16–18]. In the Szilard engine protocol, $H(\lambda_0) = H(\lambda_f) = 0$, $\rho_0 = |0\rangle\langle 0|$. We thus set $\widetilde{\gamma} = |0\rangle\langle 0|$. The D_0 term then becomes $\ln 2$, so that $W^0_{IN} =$ penalty - optimum $= kTD_\infty - kT\ln 2$. In the methods we derive a master equation for the characteristic function $Z(\xi) = \langle e^{\xi w}\rangle$ of the work distribution function $P(w)$. This new master equation allows us to calculate efficiently the characteristic function, the work distribution itself (Figure 4), and a bound for D_∞ (see the Methods).

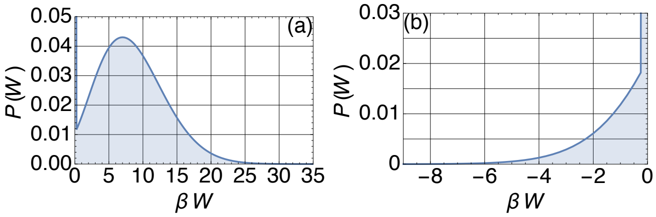

Figure 4: Work distributions calculated analytically for the forward (a) and reverse (b) process on an electron box. The two levels initially have the same energy, and one of them is lifted linearly up to $50k_BT$ and back to 0. The values of the zero-energy tunnelling rate Γ_0 and the operation time τ are set such that $\Gamma_0\tau/\varepsilon_c = k_BT$, where ε_c is the relaxation time of the metallic electrode (charge reservoir).

As described in the *trajectories* section, these results also apply if the evolution includes unitaries that create energy coherences. One might think that coherences will always worsen the worst-case work or its probability. As a counter-example according to this trajectory model, suppose $H(\lambda_0) = 0$; $\rho_0 = 1/3|0\rangle\langle 0| + 2/3|1\rangle\langle 1|$, and $H(\lambda_f) = \delta E|i\rangle\langle i|$. If the energy eigenstates stay the same throughout such that $|i\rangle = |1\rangle$ the worst-case work is δE and it has probability 2/3 (even if the shift is done quickly). If instead the Hamiltonian eigenstates change such that $|0\rangle \to |+\rangle$, and $|1\rangle \to |i\rangle = |-\rangle$ then the worst-case work is still δE corresponding to outcome $|-\rangle$ of the final energy measurement. However the probability of this can be as low as 1/2 (if H is changed suddenly $p(|-\rangle) = Tr(\rho_0|-\rangle\langle -|) = 1/2$). This shows that the probability of the worst case can actually be improved (lowered) by coherence due to suddenly changing then Hamiltonian, though at the cost of randomising the work distribution.

In the Methods we go further and consider a smaller subset of trajectories, cutting away also trajectories that start in a likely initial state but nevertheless have low probability. We describe the continuous time limit, and the electron box scenario in detail.

Summary and outlook—We showed that in any protocol with a time-varying Hamiltonian and thermalizations, the worst-case work takes the form of "penalty - optimum". The model we used could be generalised in various ways, including non-Markovian baths and baths that decohere in other bases than the energy basis. It is also important to find more bounds for the penalty term in terms of controllable parameters.

Note added: Similar results were obtained independently by Salek and Wiesner, using a different set-up and different starting assumptions, in: *Fluctuations in Single-Shot ϵ-deterministic Work Extractions.*

Acknowledgements—We are grateful to comments on a draft by Dario Egloff. We acknowledge funding from the EPSRC (UK), the Templeton Foundation, the Leverhulme Trust, the Oxford Martin School, the National Research Foundation (Singapore), the EU collaborative project TherMiQ (Grant agreement No. 618074), the Ministry of Education (Singapore), NSF grant PHY-0803371, an IQIM Fellowship, and a Virginia Gilloon Fellowship. M.-S.C acknowledges the support of the BK21 Plus Project from the Ministry of Education of Korea. The Institute for Quantum Information and Matter (IQIM) is an NSF Physics Frontiers Center supported by the Gordon and Betty Moore Foundation.

[1] G. E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3):2721–2726, September 1999.

[2] C. Jarzynski. Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters*, 78(14):2690–2693, April 1997.

[3] O. C. O. Dahlsten, R. Renner, E. Rieper, and V. Vedral. Inadequacy of von Neumann entropy for characterizing extractable work. *New Journal of Physics*, 13(5):053015, May 2011.

[4] L. del Rio, J. Åberg, R. Renner, O. C. O. Dahlsten, and V. Vedral. The thermodynamic meaning of negative entropy. *Nature*, 474(7349):61–3, June 2011.

[5] M. Horodecki and J. Oppenheim. Fundamental limitations for quantum and nanoscale thermodynamics. *Nature Communications*, 4:2059, January 2013.

[6] J. Åberg. Truly work-like work extraction via a single-shot analysis. *Nature communications*, 4:1925, January 2013.

[7] D. Egloff, O. C. O. Dahlsten, R. Renner, and V. Vedral. Laws of thermodynamics beyond the von Neumann regime. *arXiv e-print:1207.0434*, July 2012.

[8] P. Faist, F. Dupuis, J. Oppenheim, and R. Renner. A Quantitative Landauer's Principle. *arXiv e-print:1211.1037*, November 2012.

[9] Oscar C. O. Dahlsten. Non-equilibrium statistical mechanics inspired by modern information theory. *Entropy*, 15(12):5346–5361, 2013.

[10] Fernando Brandão, Micha? Horodecki, Nelly Ng, Jonathan Oppenheim, and Stephanie Wehner. The second laws of quantum thermodynamics. *Proceedings of the National Academy of Sciences*, 112(11):3275–3279, 2015.

[11] N. Yunger Halpern and J. M. Renes. Beyond heat baths: Generalized resource theories for small-scale thermodynamics. *ArXiv e-prints:1409.3998*, September 2014.

[12] C. Browne, A. J. P. Garner, O. C. O. Dahlsten, and V. Vedral. Guaranteed energy-efficient bit reset in finite time. *Phys. Rev. Lett.*, 113:100603, Sep 2014.

[13] R. Renner. *Security of Quantum Key Distribution.* PhD thesis, ETH Zürich, December 2005.

[14] R. Renner and S. Wolf. Smooth renyi entropy and applications. In *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, pages 232–232. IEEE, 2004.

[15] N. Yunger Halpern, A. J. P. Garner, O. C. O. Dahlsten, and V. Vedral. Unification of fluctuation theorems and one-shot statistical mechanics. *ArXiv e-prints:1409.3878*, September 2014.

[16] Jonne V. Koski, Ville F. Maisi, Jukka P. Pekola, and Dmitri V. Averin. Experimental realization of a szilard engine with a single electron. *Proceedings of the National Academy of Sciences*, 111(38):13786–13789, 2014.

[17] J. V. Koski, T. Sagawa, O-P. Saira, Y. Yoon, A. Kutvonen, P. Solinas, M. Mottonen, T. Ala-Nissila, and J. P. Pekola. Distribution of entropy production in a single-electron box. *Nat Phys*, 9(10):644–648, 10 2013.

[18] O.-P. Saira, Y. Yoon, T. Tanttu, M. Möttönen, D. V. Averin, and J. P. Pekola. Test of the jarzynski and crooks fluctuation relations in an electronic system. *Phys. Rev. Lett.*, 109:180601, Oct 2012.

[19] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information.* Cambridge University Press, 2010.

[20] Nilanjana Datta. Min-and max-relative entropies and a new entanglement monotone. *IEEE T. Inform. Theory*, 55(6):2816–2826, 2009.

[21] H. T. Quan and H. Dong. Quantum Crooks fluctuation theorem and quantum Jarzynski equality in the presence of a reservoir. *arXiv e-print*, page 6, December 2008.

[22] Yûto Murashita, Ken Funo, and Masahito Ueda. Nonequilibrium equalities in absolutely irreversible processes. *Phys. Rev. E*, 90:042110, Oct 2014.

[23] Tameem Albash, Sergio Boixo, Daniel A Lidar, and Paolo Zanardi. Quantum adiabatic markovian master equations. *New Journal of Physics*, 14(12):123016, December 2012.

[24] G.-L. Ingold and Yu. V. Nazarov. Charge tunneling rates in ultrasmall junctions. In *Single Charge Tunneling: Coulomb Blockade Phenomena in Nanostructures.* Plenum, 1992.

[25] M. Amman, R. Wilkins, E. Ben-Jabob, P. D. Maker, and R. C. Jaklevic. Analytic solution for the current-voltage characteristic of two mesoscopic tunnel junctions coupled in series. *Phys. Rev. B*, 43(1):1146, 1991.

Appendix A: Properties of $\widetilde{\gamma}$ and associated protocol

For a given initial state $\rho = \sum p_i |i\rangle\langle i|$ and initial energy eigenvalues E_i, the associated thermal state is defined as $\widetilde{\gamma} = \sum_i e^{-\beta\widetilde{E_i}}/\widetilde{Z}|i\rangle\langle i|$, where $\widetilde{E_i} = E_i$ for $|i\rangle \in \mathcal{E}_{\rm IN}$, but for $|i\rangle \in \mathcal{E}_{\rm OUT}$, $\widetilde{E_i}$ is chosen such that $e^{-\beta\widetilde{E_i}}/\widetilde{Z} = p_i$. Physically, this implies replacing the energy levels with small occupation probability p_i by much higher energy levels such that their thermal occupation probability is as small as p_i. The Hamiltonian associated with $\widetilde{\gamma}$ is accordingly $\widetilde{H} := \sum_{\rm IN} E_i|i\rangle\langle i| + \sum_{\rm OUT} \widetilde{E_i}|i\rangle\langle i|$. The normalising factor is $\widetilde{Z} = \sum_{|i\rangle} e^{-\beta\widetilde{E_i}}$. These definitions imply that

$$\widetilde{Z} = \frac{\sum_{|i\rangle\in\mathcal{E}_{\rm IN}} e^{-\beta E_i}}{1 - p(OUT)}. \tag{A1}$$

Apart from the given actual protocol, we also design a $\sim$-protocol such that it gives the same worst-case work in the case of $\widetilde{\gamma}$ as the initial state. We define the $\sim$-protocol as beginning with $\widetilde{H}$, then lowering the OUT levels back to E_i, i.e. setting $\widetilde{H} \to H$. After that it is the same as the actual protocol. We call the $\sim$-protocol applied to $\widetilde{\gamma}$ "the $\sim$-scenario."

In the $\sim$-scenario we similarly have $\widetilde{\mathcal{T}}_{\rm IN}$ and $\widetilde{\mathcal{T}}_{\rm OUT}$, and $\widetilde{w}^0_{\rm IN}$. The following holds:

$$\widetilde{w}^0_{\rm IN} = w^0_{\rm IN}, \tag{A2}$$

i.e. the worst-case work is the same in the $\sim$-scenario as in the actual scenario, for the $\mathcal{T}_{\rm IN}$ subset of trajectories. This is because the protocol is defined such that the added initial step in the $\sim$-scenario only involves the OUT levels. The set of possible work values are the same in $\mathcal{T}_{\rm IN}$ and $\widetilde{\mathcal{T}}_{\rm IN}$.

We now make the following mild restriction on protocols allowed:

$$\widetilde{w}^0_{\rm IN} = \widetilde{w}^0. \tag{A3}$$

We say this is mild, because the trajectories $\widetilde{\mathcal{T}}_{\rm OUT}$ have an extra work *gain* relative to their sister trajectories in $\mathcal{T}_{\rm OUT}$ following from their initial lowering. This gain tends to infinity as $p(OUT) \to 0$. The restriction of equation (A3) then means that the negative infinite work from a $\widetilde{\mathcal{T}}_{\rm OUT}$ trajectory is not a worse work cost than that from a $\mathcal{T}_{\rm IN}$ trajectory. Combining Eqs.A2 and A3 gives the desired expression used in the main body:

$$w^0_{\rm IN} = \widetilde{w}^0.$$

Appendix B: Smooth relative entropy

As noted in the main body, the optimum term reduces to a relative entropy in a special case. If $H(\lambda_0) = H(\lambda_f)$, $\log\left(Z/\widetilde{Z}\right) = -\log\sum_{i\in supp(\rho_0)} e^{-\beta E_i}/Z_f = D_0(\rho_0||\gamma)$. Moreover if $H(\lambda_0) = H(\lambda_f) = 0$, $D_0(\rho_0||\gamma) = \log d - S_{\max}(\rho_0)$ (noting $\gamma = \mathbb{1}/d$). This recovers the known results from [3, 5, 6] that these are optimal in the respective cases. If $p(OUT)$ defined above is not necessarily zero, this optimal term depends on which levels are chosen to be in $\mathcal{E}_{OUT}$. If one chooses the *best* cut between IN and OUT, in the sense of minimising $\widetilde{Z}$ and thus the worst-case work, the optimal term becomes in those cases $D^\epsilon_0(\rho_0||\gamma) := \min D_0(\rho'||\gamma)$ such that $d(\rho_0, \rho') \leq \epsilon$ where d is the trace distance (this is called the *smooth* relative entropy). The interpretation is that the optimal worst-case work allowing for an error tolerance of $\epsilon = p(OUT)$ is $kTD^\epsilon_0(\rho_0||\gamma)$, consistent with [3, 5, 6].

Appendix C: Cutting the work-tail, as well as the state-tail

There can actually be (sets of) trajectories which are unlikely even if the initial state of the trajectory is likely, as the hopping probability may be low. For example if one lifts one level towards a very high value whilst thermalizing, there is one trajectory corresponding to staying in that level throughout, which would then be the one that gives the worst-case work. However if this is very unlikely one would wish to ignore such a trajectory when stating the worst-case work. In this section we show a way to do that, by not only cutting off a part of the initial state as previously, but also a part of the work distribution. This gives a different penalty term—lower in general—in the equality for the worst-case work.

Proof overview—We shall again take the initial density matrix to have the form $\rho_0 = \sum_{i=1}^d p_i|i_0, \lambda_0\rangle\langle i_0, \lambda_0|$, not necessarily a thermal state. Then a sequence of Hamiltonian changes and thermalizations as described above is applied. This induces some work probability distribution and some worst-case work for the trajectories of interest.

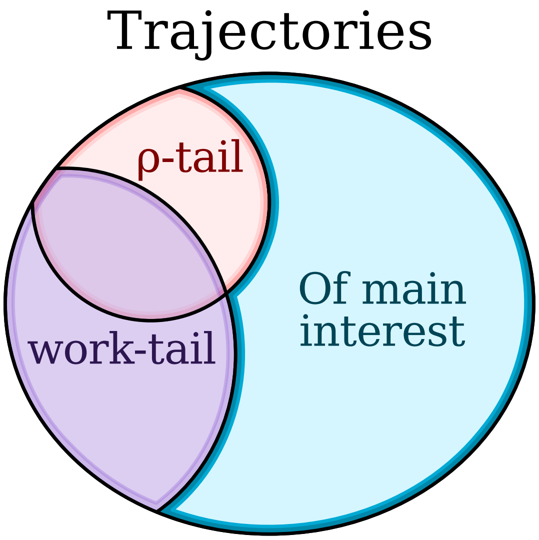

Figure 5: Depiction of the trajectories of interest. We shall ignore trajectories that have undesirable, very unlikely work values (that are in the work-tail) and trajectories that start in very unlikely energy eigenstates (that start in the ρ-tail).

The argument is split in two. First, we define a set of trajectories of interest: Some trajectories are unlikely enough to be ignorable. We derive the worst-case work for that set. Next, we consider the probability that some trajectory is in that set. Combining these two parts gives our new equality for worst-case work.

The set of trajectories of interest—We wish to ignore unlikely trajectories. We identify a set of trajectories of interest, defined as excluding trajectories of two types:

1. ρ_0-tail trajectories: These are those which are called $\mathcal{T}_{IN}$ above, i.e. trajectories which start in $\mathcal{E}_{IN}$. We now call them ρ_0-tail trajectories as using IN risks generating confusion because of the second type of cut we shall make on the set of trajectories.

2. Work-tail trajectories: We also ignore trajectories associated with the worst work values, if those values are sufficiently improbable. This ignoring amounts to cutting off the worst-case tail of the work probability distribution. To simplify the proof, we define this tail in terms of the work probability distribution of the fictional thermal state $\widetilde{\gamma}$. By "the work-tail," we mean the set of trajectories associated with the following work values w: If the initial state is $\widetilde{\gamma}$, there is an associated work probability distribution $\widetilde{p}_{\rm fwd}(w)$ for the given protocol, and an associated worst-case work $\widetilde{w}^\epsilon$. The work tail trajectories are by definition those with work cost $w > \widetilde{w}^\epsilon$. Since the actual initial state ρ_0 may differ from $\widetilde{\gamma}$, the probability that some trajectory begins in the work tail does not necessarily equal ϵ.

These sets are depicted in Fig. 5. We shall call the worst-case work in the set of interest $w^0_{IN,IN}$

The worst-case work in that set—We now derive the worst-case work in that set of trajectories, i.e. we maximise the work cost w over that set of interest. We shall for the first part draw inspiration from an argument in [15] concerning scenarios where Crooks' Theorem holds. Take the initial state of the forwards process to be $\rho_0 = \widetilde{\gamma}$; and the initial state of the reverse process as $\gamma = e^{-\beta H(\lambda_f)}/Z_f$.

Maximize Crooks' Theorem over the support of $p_{\rm fwd}(w)$ [15]:

$$\max \frac{\widetilde{p}_{\rm fwd}(w)}{p_{\rm rev}(-w)} = \max \frac{Z}{\widetilde{Z}} e^{\beta w^0}.$$

The RHS is monotonic in w, so maximizing the RHS over the support of $p_{\rm fwd}(w)$ leads to the maximum w-value w^0. Taking the logarithm and recalling the D_∞ definition yields [15],

$$\beta w^0 = D_\infty(\widetilde{p}_{\rm fwd}(w)||p_{\rm rev}(-w)) - \log\left(\frac{Z}{\widetilde{Z}}\right).$$

Now, we cut off the work tail by defining a cut-off probability distribution $\widetilde{p}^\epsilon_{\rm fwd}(w) := 0$, if $w \leq \widetilde{w}^\epsilon$ and $\frac{\widetilde{p}_{\rm fwd}(w)}{1-\epsilon}$, otherwise wherein $\widetilde{w}^\epsilon$ denotes the work guaranteed up to probability ϵ if $\widetilde{\gamma}$ is the initial state. [Dividing by $(1-\epsilon)$ normalizes the distribution.] For work values outside the work tail, Crooks' Theorem can be reformulated as

$$\frac{\widetilde{p}^\epsilon_{\rm fwd}(w)}{p_{\rm rev}(-w)}(1-\epsilon) = \frac{Z}{\widetilde{Z}} e^{\beta w}.$$

Since the RHS is monotonic,

$$\max \frac{\widetilde{p}^{\epsilon}_{\text{fwd}}(w)}{\widetilde{p}_{\text{rev}}(-w)}(1-\epsilon) = \left(\frac{Z}{\widetilde{\widetilde{Z}}}\right) e^{\beta \widetilde{w}^{\epsilon}},$$

wherein the maximization is over the support of $\widetilde{p}^{\epsilon}_{\text{fwd}}$. Taking the logarithm and rearranging yields

$$\beta \widetilde{w}^{\epsilon} = D_{\infty}(\widetilde{p}^{\epsilon}_{\text{fwd}}(w) \| \widetilde{p}_{\text{rev}}(-w)) + \log(1-\epsilon) - \log\left(\frac{Z}{\widetilde{\widetilde{Z}}}\right).$$

The LHS is the worst-case work in the set of trajectories of interest.

Probability that a trajectory is in the set of interest—The trajectories of interest are effectively the possible trajectories. To make precise what is meant by "effective," we bound the probability of not being in that set.

Consider a trajectory followed by a system initialized to ρ_0. The probability that the trajectory lies outside the set of interest is bounded by $p(\rho_0-\text{tail}) + p(\text{work}-\text{tail})$, as shown in Fig. 5. $p(\rho_0-\text{tail})$, defined via ρ_0 and the choice of effective support, is specified by input parameters. $p(\text{work}-\text{tail})$ denotes the probability that the trajectory is in the set associated with a worse work cost than $\widetilde{w}^{\epsilon}$ (the work guaranteed up to probability ϵ not to be exceeded, if the initial state is $\widetilde{\gamma}$). $p(\text{work}-\text{tail})$ does not necessarily equal ϵ for an arbitrary ρ_0. As $p(\text{work}-\text{tail})$ is not an input parameter, we wish to bound it with input parameters.

Let us drop the subscript "fwd" and refer simply to $p(w)$. The weight $p(w > x)$ in the actual work tail with ρ_0 cannot differ arbitrarily from the weight $\widetilde{p}(w > x)$ in the work tail associated with $\widetilde{\gamma}$:

$$|p(w > x) - \widetilde{p}(w > x)| \leq d(p(w), \widetilde{p}(w)).$$

This bound follows from the definition of the variation distance d, which equals the trace distance between diagonal states.[1]

The variation distance d is contractive under stochastic matrices, because the trace distance is contractive under completely positive trace-preserving (CPTP) maps. We note that the work distribution is the result of a stochastic matrix acting on the probability distribution over initial energy eigenstates. Let us now in this paragraph for convenience use Dirac notation for classical probability vectors, representing a probability distribution $p(w)$ as $\langle w|p\rangle$. The work distribution comes from the stochastic matrix $\sum_j |p_j\rangle\langle j|$ mapping a state $|\rho_0\rangle$ to a work distribution, wherein j labels projectors onto $H(\lambda_0)$ eigenstates, $|p_j\rangle$ labels the work distribution when starting with an initial state $|j\rangle$ (i.e. $p_j(w) = \langle w|p_j\rangle$), and $|\rho_0\rangle = \sum_j q_j|j\rangle$. For example, if there are two possible eigenstates, we can write $|\rho_0\rangle = q_1|1\rangle + q_2|2\rangle = (q_1\ q_2)^T$, and the resulting work distribution $p(w) = (\langle w|p_1\rangle\langle 1| + \langle w|p_2\rangle\langle 2|)|\rho_0\rangle = q_1 p_1(w) + q_2 p_2(w)$.

Thus,

$$|p(w > x) - \widetilde{p}(w > x)| \leq d(p(w), \widetilde{p}(w)) \leq d(\rho_0, \widetilde{\gamma})\ \forall x.$$

For some $x = x'$, by definition, $\widetilde{p}(w > x') = \widetilde{p}(\text{work}-\text{tail}) = \epsilon$, and $p(\text{work}-\text{tail}) := p(w > x')$. Thus

$$p(\text{work}-\text{tail}) \leq d(\rho_0, \widetilde{\gamma}) + \epsilon.$$

Main result, also cutting work tail—We conclude that the worst-case work from the trajectories of interest, $w^0_{IN,IN}$ respects

$$\beta w^0_{IN,IN} = D_{\infty}(\widetilde{p}^{\epsilon}_{\text{fwd}}(w) \| p_{\text{rev}}(-w)) + \log(1-\epsilon) - \log Z/\widetilde{\widetilde{Z}}. \tag{C1}$$

The probability that the trajectory is not in the set of interest is upper bounded by $p(\rho\text{-}tail) + p(work\text{-}tail) \leq p(\rho\text{-}tail) + d(p_i, \widetilde{\gamma}) + \epsilon$.

Appendix D: Continuous time versus discrete time

We have mainly focused on the discrete-time protocol. Experimental realizations of thermodynamic protocol are often described by a continuous master equation. Here, we show that the discrete protocol leads to a master equation in the continuum model and vice versa. In this section we restrict ourselves to scenarios without energy coherences, i.e. the discrete-classical case.

[1] See, e.g., Sec. 2 in `http://people.csail.mit.edu/costis/6896sp11/lec3s.pdf`.

1. From discrete to continuous

We consider a discrete sequence of times, $t_m = t_0 + m\,dt$ $(m = 0, 1, 2\cdots)$, and the sequence $\lambda_m \equiv \lambda(t_m)$ of values of the external parameter. As the waiting time decreases $(dt \to 0)$, the transition probability $p(|i, \lambda(t), t\rangle \to |j, \lambda(t + dt), t + dt\rangle)$ due to thermalization should vanish. To first order, it behaves as

$$p(|i, \lambda(t), t\rangle \to |j, \lambda(t + dt), t + dt\rangle) \approx \delta_{ij} + \Gamma_{i\to j}(t)dt + \mathcal{O}(dt^2). \tag{D1}$$

The transition rate $\Gamma_{i\to j}(t)$ is a possibly complicated function of instantaneous energy levels $E(|i, \lambda(t), t\rangle)$. However, the transition rates inherit the condition

$$\frac{\Gamma_{i\to j}(t)}{\Gamma_{j\to i}(t)} = e^{-\beta[E(|j,\lambda(t),t\rangle)-E(|i,\lambda(t),t\rangle)]} \tag{D2}$$

from detailed balance and the condition

$$\sum_j \Gamma_{i\to j}(t) = 0 \tag{D3}$$

from probability conservation. The occupation probability is

$$\begin{aligned} p(|j, \lambda(t + dt), t + dt\rangle) &= \sum_i p(i, |\lambda(t), t\rangle)p(|i, \lambda(t), t\rangle \to |j, \lambda(t + dt), t + dt\rangle) \\ &\approx p(|j, \lambda(t), t\rangle) + \sum_i p(i, |\lambda(t), t\rangle)\Gamma_{i\to j}(t)\,dt - \sum_i p(j, |\lambda(t), t\rangle)\Gamma_{j\to i}(t)\,dt. \end{aligned}$$

If the occupation probability is a smooth function of time, the master equation

$$\frac{d}{dt}p(|j, \lambda(t + dt), t + dt\rangle) = \sum_i p(i, |\lambda(t), t\rangle)\Gamma_{i\to j}(t) - \sum_i p(j, |\lambda(t), t\rangle)\Gamma_{j\to i}(t) \tag{D4}$$

follows. The equivalence is further illustrated in Appendix E in the example of an electron box.

2. From continuous to discrete

Going in the other direction, we now show explicitly how the discrete-time model can be derived from a physical master equation. Consider a two-level system that has a state $|0\rangle$, kept at zero energy, and a state $|1\rangle$ whose energy $\hbar\omega(t)$ changes. The Hamiltonian is $H(t) = \hbar\omega(t)|1\rangle\langle 1|$, and the system interacts with a temperature-T heat bath. In [23], a master equation for the density matrix $\rho(t)$ was derived for a such system. In the present case, the master equation is

$$\begin{aligned} \dot\rho(t) &= -i[H(t), \rho(t)] + \mathcal{L}(t)\rho(t) && \text{(D5)} \\ \mathcal{L}(t)\rho &= \Gamma d(\omega(t)) \left([n_{\rm th}(\omega(t)) + 1] \{[\sigma_-, \rho(t)\sigma_+] + h.c.\} + n_{\rm th}(\omega(t)) \{[\sigma_+, \rho(t)\sigma_-] + h.c.\}\right). && \text{(D6)} \end{aligned}$$

The heat bath's thermal photon number $n_{\rm th}(\omega) = (e^{\beta\hbar\omega} - 1)^{-1}$ depends on time because the upper level shifts. $d(\omega)$ is the dimensionless heat-bath density of states; Γ denotes a rate assumed to be constant; $\sigma_- = |0\rangle\langle 1|$ denotes the usual lowering operator; and $\sigma_+ = \sigma_-^\dagger$. Equation (D5) has the form of the usual Lindblad master equation, but the Lindblad operator depends on time. The dependence arises only from the level spacing's time dependence. The Hamiltonian part contains the Lamb shift.

In the derivation of Eq. (D5) one assumes, as usual, weak coupling to the heat bath, the Markovian approximation, and the rotating-wave approximation. One also assumes that the adiabatic approximation holds, i.e. the system always remains in its time-local energy eigenstates when the interaction with the heat bath is ignored. This condition is always fullfilled under the assumption of vanishing energy coherences at all times that we made in this section. Indeed, the part of (D5) pertaining to the diagonal elements of $\rho(t)$ can be derived without the adiabatic assumption [24].

We now consider discrete times $t_n := n\Delta t$, $n = 0, \ldots, N$, with $\omega(t)$ constant during the time intervals Δt, $\omega_n := \omega(t_n)$. Restricting ourselves to changes of the Hamiltonian that only involve its spectrum, $H(t)$ and $\mathcal{L}(t)$ are constant during a given time interval.

Consider first the Hamiltonian changes. Heisenberg's equation of motion for the system-and-bath composite implies that $\dot{\rho}(t)$ has a finite jump when the Hamiltonian has a finite jump. Therefore, $\rho(t)$ is continuous when the Hamiltonian has a finite jump. Hence for finite Hamiltonian changes during a time δt, the system-and-bath composite's density matrix is unchanged in the limit as $\delta t \to 0$. Hence the system's reduced density matrix is unchanged during the instantaneous shift of energy levels. As for the relaxation process, the initial thermal state is described in terms of occupation probabilities p_n for the n-th level. The evolution during the relaxation process is given by $p(t) = e^{Tt}p(0)$, where T is a matrix that connects the diagonal matrix elements of ρ in the master equation (D5), $\dot{\rho}_{nn} = \sum_m T_{nm}\rho_{mm}$. The transition rates T_{nm} inherit detailed balance from the rates appearing in the master equation, i.e. $T_{ij} = e^{-\beta(\epsilon_i - \epsilon_j)}T_{ji}$. Expanding e^{Tt} into a power series, one realizes that for each power T^k of T detailed balance holds, i.e. $(T^k)_{ij} = e^{-\beta(\epsilon_i-\epsilon_j)}(T^k)_{ji}$ for all $k \in \mathbb{N}$, and therefore also for e^{Tt}. We thus have derived, from a physical model of a system that is coupled to a heat bath and whose energy levels are piece-wise-constant, the discrete-time model considered in the paper.

To illustrate this let us consider a two level system: Expressing $\rho(t) = p_0(t)|0\rangle\langle 0| + [1 - p_0(t)]|1\rangle\langle 1|$, we obtain a differential equation for $p_0(t)$,

$$\dot{p}_0(t) + g(\omega(t))p(t) = \frac{g(\omega(t))}{2} + \Gamma d(\omega(t)), \tag{D7}$$

wherein $g(\omega(t)) := 2\Gamma[2n_{\mathrm{th}}(\omega(t)) + 1]$. This equation has the general solution

$$p_0(t) \;=\; \left(p_0(0) - \frac{1}{2} + \Gamma\int_0^t dt_1 d(\omega(t_1))G(t_1)\right)/G(t) + \frac{1}{2}, \tag{D8}$$

wherein $G(t) := e^{\int_0^t g(\omega(t_1))dt_1}$. The integrals in Eq. (D8) can be calculated analytically:

$$p_0(t) = p_0(0)\mathrm{e}^{-2d(\omega)\coth(\frac{\beta\hbar\omega}{2})\Gamma t} + p_{0,\mathrm{th}}\left[1 - \mathrm{e}^{-2d(\omega)\coth(\frac{\beta\hbar\omega}{2})\Gamma t}\right], \tag{D9}$$

wherein $p_{0,\mathrm{th}} := 1/(e^{-\beta\hbar\omega}+1)$ denotes the ground state's thermal occupation. For large times, the memory of the initial state is lost, and the system relaxes towards thermal equilibrium. From Eq. (D9), we obtain the transition probabilities during relaxation over the time interval between $n\Delta t$ and $(n+1)\Delta t$: $p(|0_n, \omega_n\rangle \to |0_{n+1}, \omega_n\rangle) = p_0(\Delta t)|_{p_0(0)=1}$, $p(|0_n, \omega_n\rangle \to |1_{n+1}, \omega_n\rangle) = 1 - p(|0_n, \omega_n\rangle \to |0_{n+1}, \omega_n\rangle)$, $p(|1_n, \omega_n\rangle \to |0_{n+1}, \omega_n\rangle) = p_0(\Delta t)|_{p_0(0)=0}$, and $p(|1_n, \omega_n\rangle \to |1_{n+1}, \omega_n\rangle) = 1 - p(|1_n, \omega_n\rangle \to |0_{n+1}, \omega_n\rangle)$. These transition probabilities obey detailed balance. As they remain unchanged by the inclusion of an instantaneous Hamiltonian change at the end of each time interval, we have $p(|i_n, \omega_n\rangle \to |j_{n+1}, \omega_n\rangle) = p(|i_n, \omega_n\rangle \to |j_{n+1}, \omega_{n+1}\rangle)$ for $i, j \in \{0, 1\}$.

Appendix E: Application to solid-state system: Electron box

To demonstrate the physical relevance of our results, we take a realistic example, the so-called electron box, and apply our results to it. We first derive a time-local master equation for the level-occupation probabilities in Appendix E 1. As shown in Appendix D, it is equivalent to the discrete-time trajectory model discussed in the main text. Then the work distribution functions are analyzed numerically in Appendix E 2 and analytically in Appendix E 3. Finally, we provide an upper bound of the penalty term D_∞, which reveals the direct physical relevance of our results.

1. Theoretical model and its justification

We consider the type of system in [16–18]. Following a semiclassical theory (known as "the orthodox theory") such as in [25], we derive a master equation and illustrate the work fluctuations. While a more complete quantum description is possible [e.g., 24], the semiclassical approach is useful for interpreting and identifying work and heat, which are often ambiguous.

The system (Fig. 6) consists of a large metallic electrode R that serves as a charge reservoir, a small metallic island (or quantum dot) D, and a gate electrode. The island D is coupled only capacitively to the gate electrode but couples to the reservoir R capacitively and via tunnelling. The Hamiltonian has four parts: $H = H_R + H_D + H_C + H_T$. The first two terms,

$$H_R = \sum_k \varepsilon_k c_k^\dagger c_k \quad \text{and} \quad H_D = \sum_q \varepsilon_q d_q^\dagger d_q, \tag{E1}$$

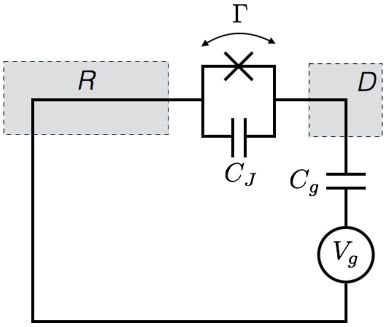

Figure 6: A schematic of an electron box.

describe the non-interacting parts of the electrode R and the island D. Here, $c_k^\dagger$ ($d_q^\dagger$) creates an electron with momentum $\hbar k$ ($\hbar q$) and energy ε_k (ε_q). The single-particle dispersions ε_k and ε_q form continua of energy levels. H_C signifies the Coulomb interaction among electrons confined in the island. Describing it within the capacitor model is sufficient:

$$H_C = \frac{Q_J^2}{2C_J} + \frac{Q_g^2}{2C_g}, \tag{E2}$$

wherein C_J and C_g denote the junction and gate capacitances, and Q_J and Q_g are equilibrium charges stored on them. One can find that

$$Q_J = C(V_g - Ne/C_g)\,, \tag{E3a}$$
$$Q_g = C(V_g + Ne/C_J)\,, \tag{E3b}$$

wherein $C := C_g C_J/(C_g + C_J)$ is the system's effective capacitance and $N = \sum_k d_k^\dagger d_k$ is the number of *excess* electrons on the island D. H_C can thus be rewritten as

$$H_C = E_C N^2 + \frac{1}{2} C V_g^2, \tag{E4}$$

wherein $E_C := e^2/2(C_g + C_J)$ is the single-electron *charging energy*, one of the largest energy scales of the system. Finally, the tunnelling of electrons between R and D is described by

$$H_T = \sum_{kq} t_{kq} c_k^\dagger d_q + h.c., \tag{E5}$$

wherein t_{kq} is the tunnelling amplitude. For common metals, which have wide conduction bands, $t_{kq} = t_d$ is independent of the momenta (or energy).

We are primarily interested in the macroscopic variable N but not in the microscopic degrees of freedom c_k and d_q, whose dynamics is typically much faster. One can thus integrate out c_k and d_q to get the effective Hamiltonian expressed only in terms of N. In the semiclassical approach, this can be achieved by considering the energy that an electron gains by tunnelling.

Suppose that an electron tunnels into the island D from the reservoir R. This will change the charge $Q_J \to Q_J - e$ and the excess number of electrons $N \to N + 1$. This new charge configuration, right after the tunnelling, is redistributed quickly to a new equilibrium configuration

$$Q_J' = C[V_g - (N+1)e/C_g] \tag{E6a}$$
$$Q_g' = C[V_g + (N+1)e/C_J] \tag{E6b}$$

by the gate voltage source. The voltage source moves the amount of charge

$$\Delta Q := Q_J' - (Q_j - e) = eC_g/(C_g + C_J) \tag{E7}$$

through the transmission line from the junction interface to the gate capacitor by doing the amount

$$W = V_g \Delta Q = eV_g C_g/(C_g + C_J) \tag{E8}$$

of work on the system. Therefore, the electron's overall energy gain ΔE is given by the work W minus the change in the electrostatic energy:

$$\Delta E = E_C \left[2C_g V_g/e - (2N+1)\right] . \tag{E9}$$

As this energy gain comes from the transition $N \to N+1$, the effective Hamiltonian for the macroscopic variable N can be regarded as

$$H_{\text{eff}} = E_C(N^2 - 2NN_g), \tag{E10}$$

wherein $N_g := C_g V_g/e$. Recall that the second term comes from the work done on the system by the voltage source.

The remaining effect of the microscopic degrees of freedom that have been removed from the macroscopic effective model is to fluctuate N randomly. As the transition $N \to N \pm 1$ is associated with tunnelling of an electron into/from the island, the transition rate can be obtained from Fermi's Golden Rule:

$$\Gamma(\Delta E) \approx \frac{2\pi |t_d|^2 \rho_R \rho_D}{\hbar} \frac{\Delta E}{e^{\beta \Delta E} + 1} , \tag{E11}$$

wherein ρ_R and ρ_D are the density of states of R and D, respectively, and

$$\Delta E = H_{\text{eff}}(N \pm 1) - H_{\text{eff}}(N) . \tag{E12}$$

Finally, at sufficiently low temperatures ($\beta E_C \gg 1$), higher changing levels play no role, and considering the two lowest levels $N = 0$ and $N = 1$ is sufficient for $N_g \in [0, 1]$.[2] Together with Eqs. (E10) and (E11), this two-level approximation leads to the master equation

$$\dot{p}_0 = -\Gamma_+ p_0 + \Gamma_- p_1 \tag{E13a}$$

$$\dot{p}_1 = -\Gamma_- p_1 + \Gamma_+ p_0, \tag{E13b}$$

wherein the transition rates are

$$\Gamma_\pm(t) := \Gamma(\pm\epsilon(t)) \quad \text{and} \quad \Gamma(\epsilon) := \frac{\Gamma_0 \epsilon(t)/\varepsilon_c}{e^{\beta\epsilon(t)} - 1}. \tag{E14}$$

Here, ε_c is the bath's high-frequency cutoff (i.e., $\hbar/\varepsilon_c$ is the correlation time), and Γ_0 is a constant that characterizes the strength of the coupling to the bath. Γ_0/ε_c is related to the material properties by $\Gamma_0/\varepsilon_c = 2\pi |t_d|^2 \rho_R \rho_D/\hbar$. Note that the transition rates satisfy the detailed-valance relation

$$\frac{\Gamma(+\epsilon)}{\Gamma(-\epsilon)} = e^{-\beta\epsilon}. \tag{E15}$$

The time-local master equation (E13) is equivalent to the discrete-time trajectory model (see Appendix D). Therefore, the electron box is a realistic prototype system to which our results can apply.

2. Monte Carlo simulation of the Electron Box

We performed a Monte Carlo of simulation of an erasure protocol in the electron box set-up. Our simulation discretizes the protocol into time steps δt that are small enough to justify the linear approximation that the population of level i evolves from time step t to $t + \delta t$ according to $p_i\left(t + \delta t\right) = p_i\left(t\right) + \delta t \dot{p}_i\left(t\right)$. Using Eqs. (E13), we can write a stochastic matrix acting on the probabilities:

$$\begin{bmatrix} P_0\left(t+\delta t\right) \\ P_1\left(t+\delta t\right) \end{bmatrix} = \begin{bmatrix} 1 - \Gamma_+ \delta t & \Gamma_- \delta t \\ \delta t \Gamma_+ & 1 - \Gamma_- \delta t \end{bmatrix} \begin{bmatrix} P_0\left(t\right) \\ P_1\left(t\right) \end{bmatrix} . \tag{E16}$$

For a two-level system which does not build up quantum coherences, a stochastic thermalizing matrix (which by its definition evolves all states towards the Gibbs state) has only one degree of freedom remaining once the Gibbs

[2] The model is invariant under $N_g \to N_g + 1$, and it suffices to regard $N_g \in [0, 1]$.

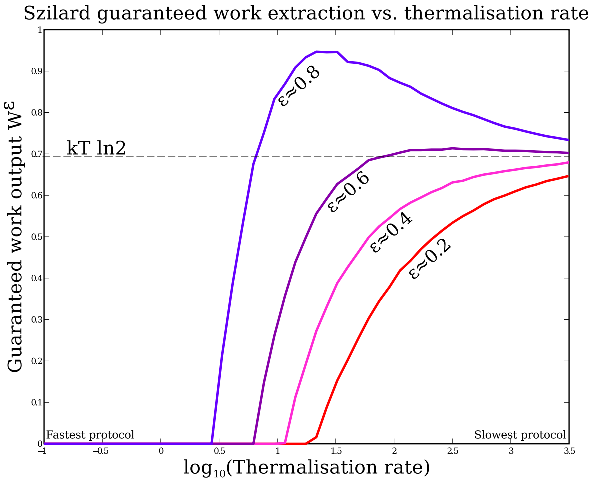

Figure 7: Work guaranteed to be extracted from a Szilárd engine up to probability ϵ: w^{ϵ}. A Monte Carlo simulation was used to predict the work from the single-electron–box. w^{ϵ} approaches $kT\ln 2$ as a function of the protocol's speed. For smaller ϵ, w^{ϵ} approaches from below; and for higher, from above.

state has been chosen: the speed of a thermalization matrix. This means that all models of two-level thermalizations for a given Gibbs state are equivalent. For our simulation we pick the conceptually straightforward *partial swap*, in which with some probability p_{sw} the current state of the system is exchanged with the Gibbs state, and otherwise it is unchanged: $M_{\mathrm{swap}} = (1 - p_{\mathrm{sw}})\mathbb{1} + p_{\mathrm{sw}}|\mathrm{Gibbs}\rangle\langle\mathrm{ones}|$, where $|ones\rangle$ means the vector of 1's. For a Gibbs state associated with an energy level splitting ϵ, we can write this explicitly as:

$$M_{\mathrm{swap}} = \begin{bmatrix} 1 - \dfrac{p_{\mathrm{sw}}\exp\left(-\beta\epsilon\right)}{1+\exp\left(-\beta\epsilon\right)} & \dfrac{p_{\mathrm{sw}}}{1+\exp(-\beta\epsilon)} \\ \dfrac{p_{\mathrm{sw}}\exp\left(-\beta\epsilon\right)}{1+\exp(-\beta\epsilon)} & 1 - p_{\mathrm{sw}}\dfrac{1-\exp\left(-\beta\epsilon\right)}{1+\exp\left(-\beta\epsilon\right)} \end{bmatrix}. \tag{E17}$$

Equating Eq. (E16) with Eq. (E17), we can find the partial swap probability in terms of the physical parameters of the electron box:

$$p_{\mathrm{sw}}\left(t\right) = \frac{\Gamma_0 \delta t}{\varepsilon_c}\epsilon(t)\coth\left[\beta\epsilon(t)/2\right], \tag{E18}$$

where we have written the swap probability $p_{\mathrm{sw}}\left(t\right)$ and the energy level splitting $\epsilon\left(t\right)$ as functions of time, to stress that this swap probability changes as the protocol evolves. Note that the probability changes only as a function of an external parameter, the splitting, (as opposed to e.g., the current state) and so Crooks' Theorem is still applicable to thermalizations of this type).

In our Monte Carlo simulation in Fig. 7, we randomly generate trajectories by picking a random initial microstate according to the initial state probability distribution, and then evolve the system by small steps, testing at each step if a swap should occur (with probability p_{sw}), and if it does, we replacing the state with a new micro-state randomly chosen from the Gibbs state associated with the current Hamiltonian. By recording which microstate is occupied when the energy level is raised, we calculate the work cost associated with a particular trajectory. Repeated runs of the simulation allow us to build up a work distribution, to which the results in this paper can be applied.

3. Analytic expression for the work distribution

The work distribution function for an electron box can also be obtained explicitly from the master equation.

The trajectory $\sigma(t) \in \{0, 1\}$ of the system is piece-wise constant, jumping discontinuously from one energy level to another at some random instants t_j $(j = 1, 2, \cdots)$. Therefore, it is specified uniquely by the initial condition σ_0, the number J of *jumps* and the corresponding instants t_j $(j = 1, 2, \cdots, J)$. Then the probability distribution function for the trajectory is given by

$$P_J(t_1,\cdots,t_J;\sigma_0)=\prod_{j=1}^{J}\Gamma((-1)^{\sigma_0+j+1}\epsilon(t_j))\exp\left[-S_J(t_1,\cdots,t_J;\sigma_0)\right] \tag{E19}$$

where the *effective action* associated with a given trajectory has been defined by

$$S_J(t_1,\cdots,t_J;\sigma_0)=\sum_{j=1}^{J+1}\int_{t_{j-1}}^{t_j}ds\,\Gamma((-1)^{\sigma_0+j+1}\epsilon(s)) \tag{E20}$$

and it is implied that $t_0=0$ and $t_{J+1}=\tau$. It is straightforward to check the normalization

$$P_0(\sigma_0)+\sum_{J=1}^{\infty}\prod_{j=1}^{J}\int_{t_{j-1}}^{\tau}dt_j\,P_J(t_1,\cdots,t_J;\sigma_0)=1\,, \tag{E21}$$

where again it is implied that $t_0=0$.

The work is only done while the system is in the state $\sigma=1$, and hence the contribution to the work along the trajectory is given by

$$W_J(t_1,\cdots,t_J;\sigma_0)=\sum_{j=1}^{J}(-1)^{\sigma_0+j}\epsilon(t_j)+(\sigma_0+J\bmod 2)\epsilon_f-\sigma_0\epsilon_0 \tag{E22}$$

The work distribution function along a trajectory with J jumps reads as

$$P_J(W;\sigma_0)=\prod_{j=1}^{J}\int_{t_{j-1}}^{\tau}dt_j\,P_J(t_1,\cdots,t_J;\sigma_0)\,\delta(W-W_J(t_1,\cdots,t_J;\sigma_0)). \tag{E23}$$

The total work distribution function can be written in a series

$$P(W)=p_0e^{-S_0(0)}\delta(W)+p_1e^{-S_0(1)}\delta(W-W_c)+\sum_{J=1}^{\infty}\sum_{\sigma_0}p_{\sigma_0}P_J(W;\sigma_0) \tag{E24}$$

$P_J(W)$ has a factor of $(\Gamma_0^2e^{-\beta\epsilon})^J$ and at low temperatures, P_J is rapidly suppressed as J increases.

The expression (E24) for the work distribution is essentially a perturbative expansion in Γ_0^2 and converges very quickly for small Γ_0. For large Γ_0, however, it becomes impractical to use it for actual calculation because of its slow convergence. Therefore, it will be useful to devise a more general method and we examine the characteristic function $Z(\xi)=\langle e^{\xi w}\rangle$ of the work distribution function $P(W)$. We first consider the characteristic function $Z_\sigma(\xi)=\langle e^{\xi w}\rangle_\sigma$ conditioned that all trajectories start from a definite initial state σ_0. Regarded as a function of the operation time τ, $Z_\sigma(\xi;\tau)$ satisfies the master equation

$$\partial_\tau Z_\sigma(\lambda;\tau)=\sum_{\sigma}\left[\Gamma_{\sigma\sigma'}(\tau)+\lambda\partial_\tau\epsilon_\sigma(\tau)\delta_{\sigma\sigma'}\right]Z_{\sigma'}(\lambda;\tau) \tag{E25}$$

and the initial condition

$$Z(\xi;0)=e^{\xi\epsilon_\sigma(0)}\,. \tag{E26}$$

Compared with the original master equation (E13) for the level occupation probability, the new master equation (E25) for the characteristic function contains additional diagonal terms. The full characteristic function is then given by

$$Z(\xi)=\sum_{\sigma_0}p_{\sigma_0}Z_{\sigma_0}(\xi). \tag{E27}$$

Recall that $Z(\xi)$ contains the same information as $P(W)$. Indeed, one can calculate $P(W)$ itself and, as shown in Section E 4 below, a bound for $D_\infty(P_{\rm fwd}(W)\|P_{\rm rev}(-W))$.

Let us now show that the work distribution in Eq. (E24) satisfies the Crooks fluctuation theorem:

$$\frac{P_{\rm fwd}(W)}{P_{\rm rev}(-W)}=\frac{Z_f}{Z_0}e^{\beta W} \tag{E28}$$

where Z_0 and Z_f are the partition functions for the initial and final Hamiltonian in the forward protocol, respectively. Given a *forward ramping* $\epsilon(t)$, the *reverse ramping* $\epsilon^{\rm rev}(t)$ is defined by

$$\epsilon^{\rm rev}(t) = \epsilon(\tau - t)\,. \tag{E29}$$

In the forward protocol, consider a trajectory $\sigma(t)$ characterized by the initial condition σ_0, the number J of energy-level jumps and the jump instants t_j ($j = 1, 2, \cdots, J$). One can find a unique trajectory $\sigma^{\rm rev}(t)$ in the reverse protocol, which is defined by the initial condition

$$\sigma_0^{\rm rev} = \sigma_0 + J \pmod 2 \tag{E30}$$

and the flip instants

$$t_j^{\rm rev} = \tau - t_{J-j+1}\,. \tag{E31}$$

Note that

$$\epsilon^{\rm rev}(t_j^{\rm rev}) = \epsilon(t_{J-j+1}). \tag{E32}$$

The effective action along the reverse trajectory is the same as that along the forward trajectory [cf. (E20)]:

$$S_J^{\rm rev}(t_1^{\rm rev}, \cdots, t_J^{\rm rev}; \sigma_0^{\rm rev}) = S_J(t_1, \cdots, t_J; \sigma_0). \tag{E33}$$

Further, the work contribution along the reverse trajectory is just the negative of that along the forward trajectory [cf. (E22)]:

$$W_J^{\rm rev}(t_1^{\rm rev}, \cdots, t_J^{\rm rev}; \sigma_0^{\rm rev}) = -W_J(t_1, \cdots, t_J; \sigma_0)\,. \tag{E34}$$

These observations lead to

$$P_J^{\rm rev}(t_1^{\rm rev}, \cdots, t_J^{\rm rev}; \sigma_0^{\rm rev}) = P_J(t_1, \cdots, t_J; \sigma_0) e^{-\beta W_J(t_1, \cdots, t_J; \sigma_0)} \exp\left[(\sigma_0 + J \bmod 2)\beta\epsilon_f - \sigma_0\beta\epsilon_0\right] \tag{E35}$$

and

$$P_J^{\rm rev}(-W; \sigma_0^{\rm rev}) = P_J(W; \sigma_0) e^{-\beta W} e^{(\sigma_0 + J \bmod 2)\beta\epsilon_f - \sigma_0\beta\epsilon_0} \tag{E36}$$

It is then straightforward to prove Crooks' Theorem:

$$\begin{aligned}
P_{\rm rev}(-W) &= \frac{1}{1 + e^{-\beta\epsilon_0^{\rm rev}}} \sum_{J=0}^{\infty} \sum_{\sigma_0^{\rm rev}} e^{-\beta\sigma_0^{\rm rev}\epsilon_f^{\rm rev}} P_J^{\rm rev}(-W; \sigma_0^{\rm rev}) \\
&= \frac{1}{1 + e^{-\beta\epsilon_f}} \sum_{J=0}^{\infty} \sum_{\sigma_0} e^{-\beta(\sigma_0 + J \bmod 2)\epsilon_0} \\
&\qquad \times e^{-\beta W} e^{\beta(\sigma_0 + J \bmod 2)\epsilon_f - \sigma_0\beta\epsilon_0} P_J(W; \sigma_0) \\
&= \frac{e^{-\beta W}}{1 + e^{-\beta\epsilon_f}} \sum_{J=0}^{\infty} \sum_{\sigma_0} e^{-\sigma_0\beta\epsilon_0} P_J(W; \sigma_0) \\
&= e^{-\beta W} \frac{1 + e^{-\beta\epsilon_0}}{1 + e^{-\beta\epsilon_f}} P_{\rm fwd}(W).
\end{aligned}$$

4. Upper bound of D_∞ term

Recall the Markov inequality, for a non-negative random variable X:

$$p(X \geq a) \leq \langle X \rangle / a.$$

This is derived by noting that there cannot be too much probability of having a value much greater than the average, or else the average would have to be greater. In our case it reads

$$p(w \geq \widetilde{w}^{\epsilon}) := \epsilon \leq \langle w \rangle / \widetilde{w}^{\epsilon}.$$

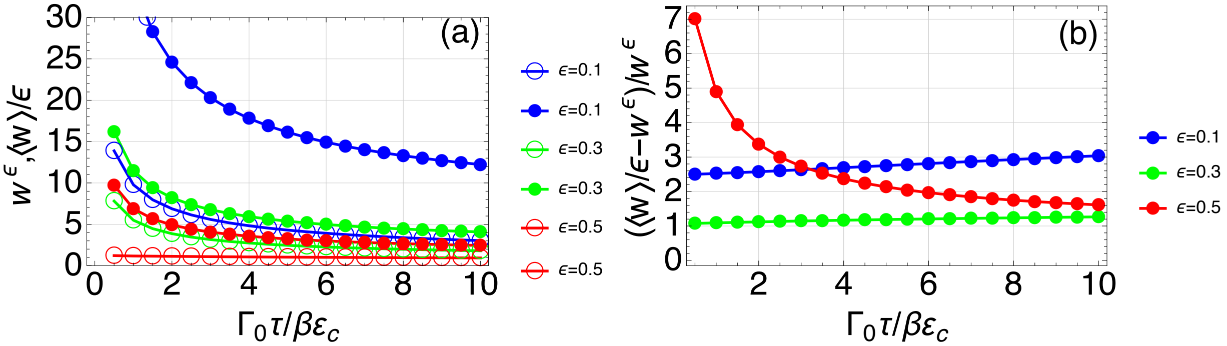

Figure 8: The $\tilde{w}^\epsilon$ and its upper bounds $\langle w\rangle/\epsilon$ (which in turn bound D_∞) for different values of ϵ. (a) Individual plots of $\tilde{w}^\epsilon$ and $\langle w\rangle/\epsilon$. (b) The relative tightness.

Thus

$$\widetilde{w}^\epsilon \leq \langle w\rangle/\epsilon.$$

Recalling the main result, and rearranging it

$$D_\infty(\widetilde{p}^\epsilon_{\mathrm{fwd}}(w)\|p_{\mathrm{rev}}(-w)) = \beta\widetilde{w}^\epsilon - \log(1-\epsilon) + \log Z/\widetilde{Z}, \tag{E37}$$

we now have

$$D_\infty(\widetilde{p}^\epsilon_{\mathrm{fwd}}(w)\|p_{\mathrm{rev}}(-w)) \leq \beta\langle w\rangle/\epsilon - \log(1-\epsilon) + \log Z/\widetilde{Z}. \tag{E38}$$

(here $\log Z/\widetilde{Z} = \log 2$). One has only to find the upper bound of $\langle w\rangle$. One can do this most easily by means of the characteristic function $\langle e^{\lambda w}\rangle$, which bounds $\langle w\rangle$ due to the convexity. This has been illustrated in Fig. 8.